AWS Data Science Project

Sponsor: Washington State University
Mentors: William Bonner & Geoff Allen
Adam Clobes, Russell Heppell, Alexander McKee, Daniel Nuon, Madison Velasquez

Washington State University
School of Electrical Engineering and Computer Science

Background

A team of researchers at WSU receive data quarterly from the Department of Revenue (DOR) and the Employment Security Department (ESD) to research and calculate Washington State business information.

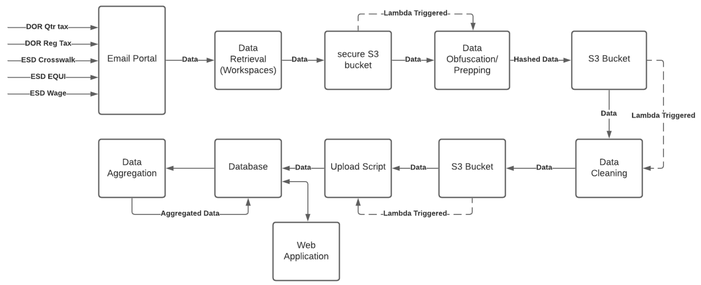
Problem

- The ESD and DOR data is stored in a protected physical location in Pullman, requiring a timely process for researchers to retrieve data.
- Data calculations/aggregations are performed every time a researcher wishes to view aggregated data.

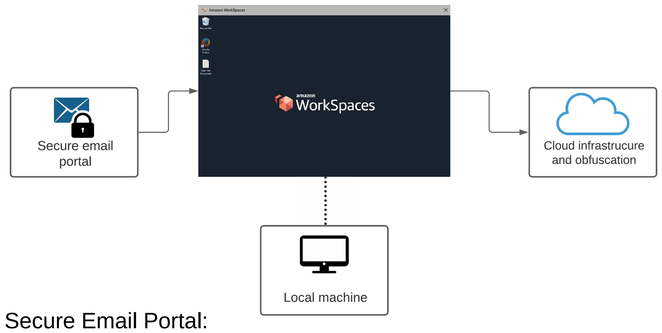
Objective

The goal of the project is to streamline the data research process, using various Amazon Web Services (AWS) to create a secure cloud-based web application with AWS cloud data storage.

Below is a diagram showing the complete flow of data throughout the system from receiving to delivery.



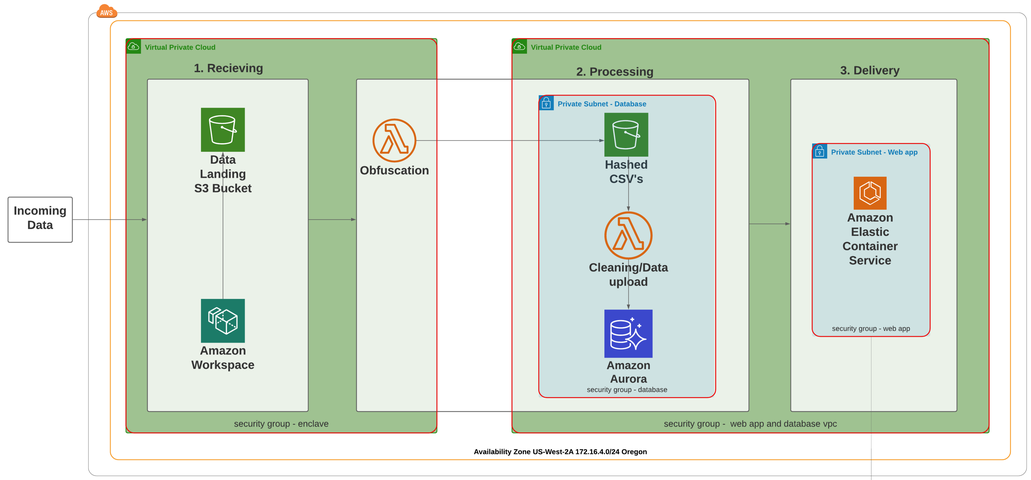
1. Receiving



- Secure Email Portal:
  - Raw data is download from a secure email portal that requires security clearance.
- AWS Workspaces:
  - Employee downloads while inside AWS Workspaces virtual machine, to insulate data from local machine.
- S3 bucket landing zone:
  - Data is loaded into S3 bucket using Identity and Access Management (IAM) privileges to ensure data cannot be moved somewhere unauthorized.

High-Level System Design

We divided our project into 3 main systems: Data receiving, processing, and delivery. The overall design can be seen below, showing the process flow using different AWS services.



Security Group Traffic Rules

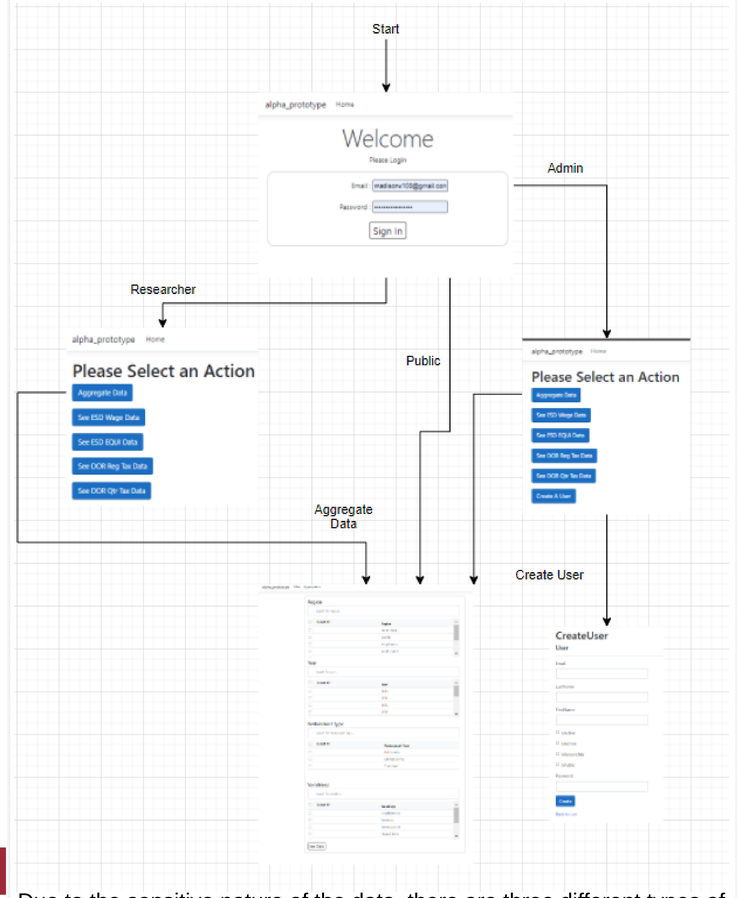
Security group (sg)	inbound	outbound
Enclave	Deny All	Deny All
Database	web app sg, (obfuscation)	webapp sg
web app	database sg, WSU IP	database sg, WSU IP
web app and database vpc	WSU IP	WSU IP

2. Processing

- Obfuscation:
  - Once file is uploaded to secure S3 Bucket, obfuscation is automatically triggered.
  - Uses industry standard sha256 encryption, ensuring sensitive data are secure.
- Cleaning:
  - Data may have trailing whitespace, lack a data column, or may not be the correct data type.
  - This process uses a python library, pandas, and knowledge of desired types to clean and prep data, so it is ready for upload to the database.
- Data upload script:
  - Data is uploaded to the database using a startup windows app inside a private server. The app reads all .csv files to context arrays and pushes them to the database.
- Data storage:
  - Achieved using an Amazon RDS PostgreSQL Server instance.
  - Two databases: Obfuscated incoming data and helper tables, Aggregated data to be populated from inside the database. Databases and their tables outlined below.

ObfuscatedData	AggregatedData
Crosswalk_File	Average_Employment_Data
ESD_EQUI	Closure_Rate_Data
ESD_Tax	Cost_of_Labor
DOR_reg_tax	Employment_Hours_Data
DOR_qtr_tax	Establishments
Geo_Keys	Min_Wage_Data
Geo_Walk	Revenue_Data
User_Login	

3. Delivery



Due to the sensitive nature of the data, there are three different types of users with different permissions.
User Types:

- Admin: Can create a user, view row level data for ESD and DOR, and view aggregated data.
- Researcher: View row level data and view aggregated data.
- Public: View aggregated data.

Glossary

Aggregation: Organizing data in an informative context.

EC2: Elastic Compute Cloud; scaling computation server that operates remotely.

Hashing: a process that converts data into unreadable text.

Obfuscation: the process of obscuring data to make it difficult to understand.

S3 Bucket: Simple Storage Service; remote storage for objects and files.

Acknowledgements

Special thanks to our mentors William Bonner and Geoff Allen, researchers Brad Gaolach and Mark Beattie, our professor Dr. Zeng, and Washington State University for supporting us in this project.

Team Aang